Applied Regression Analysis

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics, University of Washington

Session 10

Applied Regression Analysis

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics, University of Washington

Part 5:

Case Diagnostics

Lecture Outline

- Topics:
  - Outliers
  - Influence
  - Applications with Interactions

Case Diagnostics

Detecting Unusual Cases

- When using regression models to explore associations between variables, we are always very interested in whether there are individual cases that behave somewhat differently than the bulk of the data

Detecting Unusual Cases

- Some cases may be poorly described by the overall regression model
  - “Outliers”
- Some cases may be overly influential in fitting the regression model
  - “Influential cases” affect estimates
  - “Highly leveraged cases” affect statistical significance

Outliers

- "Outliers" are cases whose response is far from that predicted by the model as judged by the residual
  - Well developed for linear regression, providing you assume normally distributed data
    - – Consider how many SD a single case is from its group mean relative to the sample size of the data set
      - » The expected magnitude of the largest residual is a function of n
    - – (Lacking anything else, still probably reasonable)

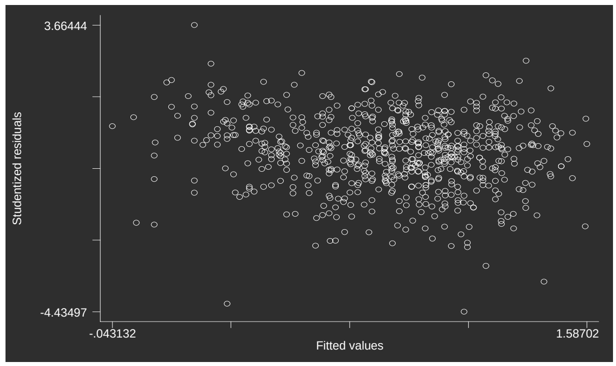
Multiple Regression Model

```
. regress logfev smoker age loght if age>=9
Number of obs =     439
Prob > F      =  0.0000
R-squared     =  0.6703
Root MSE      =  .14407

logfev |  Coef. StErr.     t   P>|t|    [95% CI]
smoker |  -.054  .0209  -2.56  0.011   -.095   -.012
   age |   .022  .0038   5.64  0.000    .014    .029
 loght |  2.870  .1301  22.06  0.000   2.614   3.125
 _cons |-11.095  .5201 -21.33  0.000 -12.117 -10.072
```

Example: FEV and Smoking

- Plot of residuals versus predicted values



Example: FEV and Smoking

- From residual plot we note extreme residuals
  – One large positive residual 3.664 standard deviations from 0
    - Based on the t distribution with 435 degrees of freedom, we would only expect 0.0139% of residuals to be this large if the log transformed FEV data were normally distributed within groups

Example: FEV and Smoking

– Large negative residuals -4.435, -4.215, and -3.593 standard deviations from 0
  - Based on the t distribution with 435 degrees of freedom, we would only expect 0.00058%, 0.00152% and 0.0182%, respectively, of studentized residuals to be this small if the log transformed FEV data were normally distributed within groups

Detecting Unusual Cases

- How large a residual is too large?
  – We are generally looking at the most extreme residuals, and thus we must account for the sample size when considering what is too extreme

Detecting Unusual Cases

- This can be thought of as a multiple comparison problem
  – Comparing the largest (smallest) residual to some threshold is equivalent to comparing every residual to that threshold
  – If we have a 5% error with each such comparison, our total error is much higher

Multiple Comparison Problem

- In frequentist reasoning, we try to ensure that our error rate is at some specified level α
  – When only making one decision, this is relatively easy

Multiple Comparison Problem

- When making multiple decisions, we must consider the "experiment-wise" error
  - Worst case scenario: An error rate of α on each decision could lead to an experiment-wise error as high as $k\alpha$
    - This would be the situation if all of our errors were "mutually exclusive"
  - If all our errors were independent of each other, our experiment-wise error is $1-(1-\alpha)^k$

Multiple Comparison Problem

Number of Comparisons	Experiment-wise Error Rate: Worst Case Scenario	Experiment-wise Error Rate: Independent Errors
1	.0500	.0500
2	.1000	.0975
3	.1500	.1426
5	.2500	.2262
10	.5000	.4013
20	1.0000	.6415
50	1.0000	.9231

Bonferroni Correction

- Assume the worst case scenario
  - When making k comparisons
    - Test individual P values against α / k, <u>OR</u>
    - Multiply P values by k and compare to α
      - (But don't get absurd: P values can never be above 1)
- Easy, but conservative

Detecting Unusual Cases

- For a level α test of outliers in a dataset containing n observations
    - Compute an individual P value for each residual based on the t distribution with n-p degrees of freedom, where p= number of regression parameters
    - Bonferroni: Compare the P value associated with the absolute value of each outlier to α / (2n)

FEV Example

- Applying the Bonferroni correction identifies four cases with extreme residuals, when we presume normally distributed residuals
  - But why do we think the FEV is lognormal within age, height, smoking groups?
  - Lack of effort would logically lead to skewed distribution of residuals

Detecting Influential Cases

- “Influential” cases are those cases which affect our inference too much
  - Such cases can affect our inference by
    - Changing the scientific estimate of association markedly from what it would be if the case were not in the data set
    - Changing the strength of statistical evidence (e.g., P value) markedly from what it would be if the case were not in the data set

Detecting Influential Cases

- Finding influential cases is conceptually quite easy
  - In turn, leave each case out and see what happens
  - There can, of course, be influential pairs (triples, etc.) of cases, but trying to detect these is hampered by the "curse of dimensionality"

Detecting Influential Cases

- In linear regression, influence of individual cases on the scientific estimates can be computed without fitting all the additional regressions
  - In other forms of regressions, "one-step" approximations are often used to assess the approximate influence of a case

Detecting Influential Cases

- Personally, I would rather separate the scientific measures of influence from the statistical measures of influence
  - Scientific: Slope when each case is deleted
  - Statistical:P value when each case is deleted

Influential Cases with Interactions

- Interactions can often appear statistically significant when some outlier is present in the data
  - Interactions are often able to make a model fit the outlier better
  - But, I am very loathe to introduce an interaction into a model just to fit an outlier
  - I examine influence of cases whenever I consider interactions

FEV Example

- We could also consider sex, age, height interactions in the FEV data set
  - We find a statistically significant interaction between sex, age, and height
  - If we leave out the two cases with the large negative residuals, there is no statistically significant association
    - I choose to not model the interaction as it is likely driven largely by those outliers

Example: SEP "Normal Ranges"

- We want to find normal ranges for somatosensory evoked potential (SEP)
  - As a first step, we want to consider important predictors of nerve conduction times
    - If any variables such as sex, age, height, race, etc. are important predictors of nerve conduction times, then it would make most sense to obtain normal ranges within such groups

Example: SEP "Normal Ranges"

- Scientifically, we might expect that height, age, and sex are related to the nerve conduction time
  - Nerve length should matter, and height is a surrogate for nerve length
  - Age might affect nerve conduction times: People slow down with age
  - Sex: Men are <u>SO</u> fragile

Example: SEP "Normal Ranges"

- Prior to looking at the data, we can also consider the possibility that interactions between these variables might be important
  - Height - age interaction?
    - Do we expect the difference in conduction times between 6 foot tall and 5 foot tall 20 year olds to be the same as the difference in conduction times between 6 foot tall and 5 foot tall 80 year olds?

Example: SEP "Normal Ranges"

- We might suspect such an interaction due to the fact that height may not be as good a surrogate for nerve length in older people
  - With age, some people tend to shrink due to osteoporosis and compression of intervertebral discs
    - It is not clear that nerve length would be altered in such a process

Example: SEP "Normal Ranges"

- Thus, in young people, differences in height probably are a better measure of nerve length than in old people
    - Tall old people probably have been tall always
    - Short old people will include some who were much taller when they were young

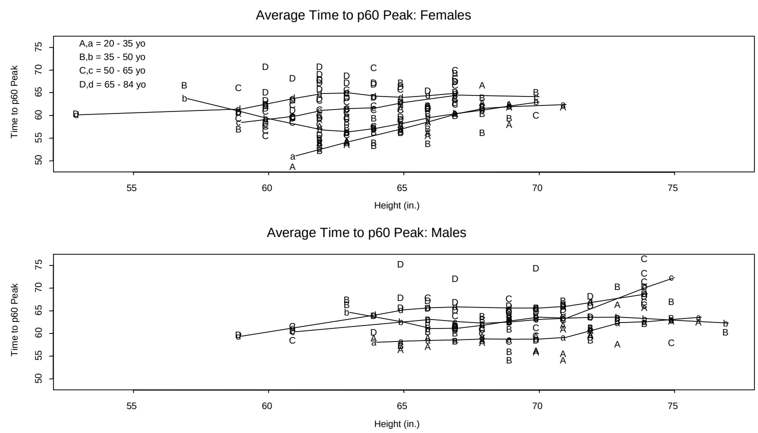
Example: SEP "Normal Ranges"

- We can also consider the possibility of three way interactions between height, age, and sex
  - Osteoporosis affects women far more than men
    - Hence, we might expect the height - age interaction to be greatest in women and not so important in men

Example: SEP "Normal Ranges"

- A two way interaction between height and age that is different between men and women defines a three way interaction between height, age, and sex

Stratified Scatterplots



Example: SEP "Normal Ranges"

- Defining a regression model with interactions
  - We must create variables to model the three way interaction term

Example: SEP "Normal Ranges"

- Furthermore, it is a <u>VERY GOOD</u> idea to include all "main effects" and "lower order interactions" in the model as well
  - "main effects": the individual variables which contribute to the interaction
  - "lower order terms": all interactions that involve some combination of the variables which contribute to the interaction

Example: SEP "Normal Ranges"

- Most often, we lack sufficient information to be able to guess what the true form of an interaction might be
  - The most popular approach is thus to consider multiplicative interactions
    - Create a new variable by merely multiplying the two (or more) interacting predictors

Example: SEP "Normal Ranges"

- Thus for this problem we could create variables
  - H.A = Height * Age
  - H.M = Height * Male
  - A.M = Age * Male
  - H.A.M = Height * Age * Male

Example: SEP "Normal Ranges"

- Interpretation of the model parameters
  - In the presence of higher order terms (powers, interactions) interpretation of parameters is not easy
    - We can no longer use "the change associated with a 1 unit difference in predictor holding other variables constant"
      - It is generally impossible to hold other variables constant when changing a covariate involved in an interaction
      - If not impossible, it is often uninteresting

Example: SEP "Normal Ranges"

Interpretation of the model in terms of the SEP height relationship within age-sex strata

Example: SEP "Normal Ranges"

$$E(p60 \mid Ht, Age, Male) = \beta_0 + \beta_1 Ht + \beta_2 Age + \beta_3 Male$$
$$+ \beta_4 H.A + \beta_5 H.M + \beta_6 A.M + \beta_7 H.A.M$$

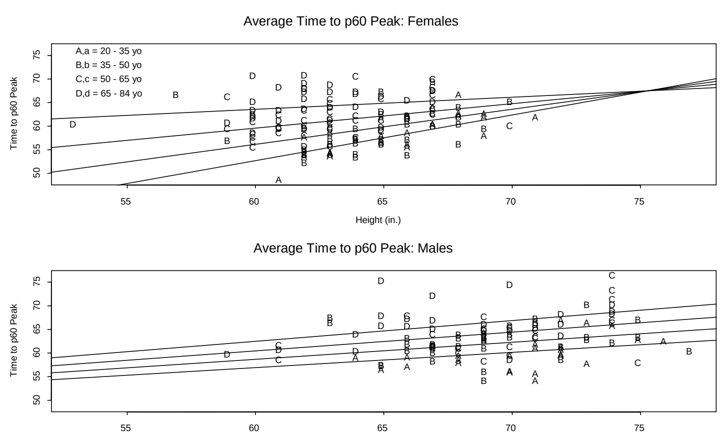
p60 – Height relationship for Age = a :

Sex	Intercept	Slope
F	$(\beta_0 + \beta_2 a)$	$(\beta_1 + \beta_4 a)$
M	$(\beta_0 + \beta_3 + (\beta_2 + \beta_6)a)$	$(\beta_1 + \beta_5 + (\beta_4 + \beta_7)a)$

Example: SEP "Normal Ranges"

- From the above, we see the importance of including the main effects and lower order terms
  - E.g., leaving out the height - sex interaction is tantamount to claiming that the p60 - height relationship among newborns is the same for the two sexes
    - (It might be, but the chance that our lines would predict the truth is very slight-- we are trying to approximate relationships in other age ranges)

Lines Predicted By Model



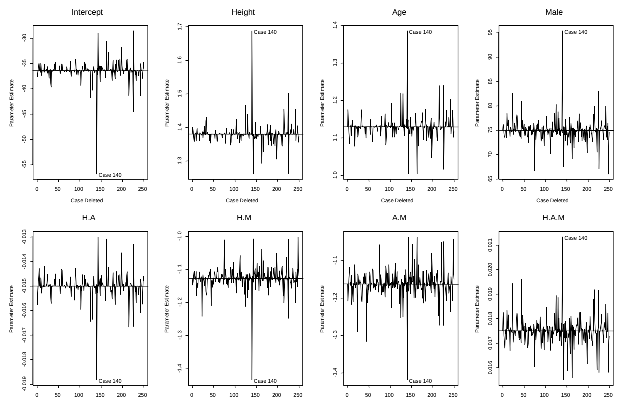
Example: SEP "Normal Ranges"

- From the inference, we find a statistically significant three way interaction
  - P= .0471

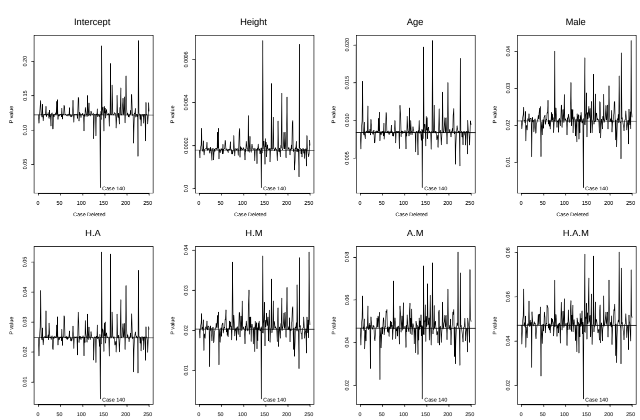
Example: SEP "Normal Ranges"

- I am now interested in ensuring that the evidence for an interaction is not based solely on a single person's observation
  - Hence, I consider 250 different regressions in which I leave out each case in turn
  - I plot the slope estimates and P values for each variable as a function of which case I left out
    - Case 0 corresponds to using the full data set

Influence on Estimated Parameters



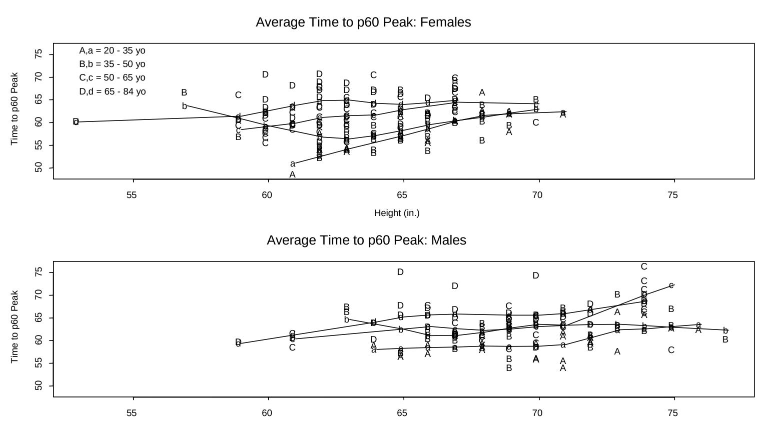
Influence on P values



Example: SEP "Normal Ranges"

- Contrary to what I was afraid of, the only influential case actually lessened the evidence of an interaction
  - When Case 140 is removed from the data, the evidence for an interaction is a larger estimate and a lower P value
  - We can examine the scatterplot to see why Case 140 might be so influential

Stratified Scatterplots



Example: SEP "Normal Ranges"

- So now what do I do with Case 140
  - From the influence diagnostics, I now feel comfortable with the fact that the data really do suggest a three way interaction

Example: SEP "Normal Ranges"

- Personally, I do <u>NOT</u> remove the case from the dataset when making my prediction intervals
    - I do not know why Case 140 is so unusual
    - It is possible that people like her are actually more prevalent in the population than my sample would suggest
      - My best guess is that she represents 0.4% of the population, so leave her in